ARTICLE DETAILS

TITLE (PROVISIONAL)	Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population.
AUTHORS	Ellender, Claire; Zahir, Syeda; Meaklim, Hailey; Joyce, Rosemarie; Cunnington, David; Swieca, John

VERSION 1 – REVIEW

REVIEWER	Scott, Hannah Flinders University
REVIEW RETURNED	04-Dec-2020

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. I applaud the authors for assessing the accuracy of consumer sleep devices in a clinical setting with clinical populations. More efforts like this are needed. My comments are largely related to the data analysis: the request for additional analyses and further explanation of the original analyses. Whilst my comments appear extensive, I do not think they would be difficult for the authors to address, yet I think they are necessary to support the interpretation of the data. Major comments: - How were lights on/off time synced across all devices with PSG? This is essential as it would impact each sleep parameter, particularly the wake parameters if the start/end of the recordings are unaligned. - The mean difference between PSG and the three devices are calculated for each sleep parameter, with the direction of the mean difference indicating whether the device under/overestimated the variables. Yet, whether the devices significantly under/overestimate the variables is yet to be tested in the current study. These analyses should be performed so the authors can conclude whether the under/overestimations are significant and substantial in magnitude, e.g. via interpretation of standardised effect sizes like cohen's d. - Further description about the Bland-Altman Plots in the 'Statistical Analyses' section would be welcome, e.g. contains the mean difference (bias), LOAs (levels of agreement), CIs of the bias and LOAs, proportional bias testing? Note the comments about proportional bias and LOA calculation below. - Regarding this statement in the Results "No significant were found between PSG and the test devices", which analyses were performed? More information about the analyses is required. - See this paper with guidelines on the conduct and reporting of device validation studies in Sleep: DOI: 10.1093/sleep/zsz254.

	Consider citing this article and whether any additions need to be made to the reporting of the current study in accordance with these guidelines. This article also provides guidance on the types of statistical comparisons that should be made, as suggested here. - The first paragraph of the Discussion section compares the accuracy of each device. Noting the comment above about the need for statistical comparison between PSG and each device, whether each consumer device significantly under/overestimates more so than another consumer device can/should be determined statistically. - An optional request: it would be interesting and clinically relevant to check whether the devices differ in performance between sleep disorders. Noting the small number of insomnia and hypersomnia patients, it would still be useful to statistically test whether the accuracy of the devices differ between OSA patients and the other two sleep disorders. - Regarding this comment in the Discussion: “A further weakness was that actigraphy was not directly compared with the consumer grade devices.” Are the authors referring to epoch-by-epoch comparisons? Either way, greater clarification on this point would be welcome. I also suggest further discussion about the limitation of not having performed epoch-by-epoch analyses. In previous studies, devices often look to perform relatively well on estimating sleep parameters, but the epoch-by-epoch analyses reveal a different story. This seems to be most common when looking at wake: SOL and WASO look ok and might not be significantly different to PSG, but the specificity is woeful. Some discussion and acknowledgement of this issue would be great. Minor comments: - It is more correct to say that consumer devices ‘estimate’ sleep rather than ‘measure’ sleep. These devices measure body acceleration, biomotion, heart rate, etc., which they then use to estimate sleep. Thus, I suggest a minor change to the article title and some terminology throughout (e.g. Abstract). - Perhaps report confidence intervals for mean averages given in the Abstract. - Was the PSG scorer an RPSGT? Only one scorer? If more than one scorer, can an indication of inter-rater reliability be provided? - App versions for the consumer devices? Were the apps updated during the study at all? - Perhaps state the bias and LOAs on the Bland Altman plots. Usually, these values are given on the right side of the plot. Additionally, please add lines of best fit with R squared and corresponding p values to assess proportional bias to each plot. - With reference to the Bland-Altman plot interpretations, I am confused by these statements: “Only 1 data point fell outside of the 95% CI”. Is the intention to signal the presence of outliers? A statistical test would be more appropriate for this, though I am not sure if this is what the authors are attempting to convey? In any case, some greater clarification about the intention of these statements and some statistics to support them are needed. - Please introduce the Figures when the TST results are first presented in text. Also, I understand that TST was the primary outcome, but providing separate plots for the secondary outcomes

	(SOL, WASO, SE) for each device would be very helpful and assist in comparison of the study findings to other studies. - Please calculate the levels of agreements in the Bland-Altman plots as ± 2 SDs from the mean difference (bias) rather than 95% CIs so that the plots are more comparable to other studies. - Regarding this statement in the Discussion: “Moving forward, these finding should indicate to developers, that some data storage is needed within sleep monitors to mitigate synchronisation failure.” I have only had experience with the Jawbone, but this device has internal storage. The issue with data loss typically arises either from a failure by the device to start and store data (i.e. there was no data recorded to sync in the first place), or the app fails to connect to the cloud-based servers to store the data during the retrieval process. I can’t comment on the two other devices, but I suggest checking to confirm whether they have internal storage. Minor typographical/grammatical errors: - Abstract: “sleep monitors to through comparison” - Patient and Public involvement: “often asked how about the accuracy” - Introduction: “grow to 8.5 billion” – need to add the dollar sign. - Introduction: “for TST measurement clinical sleep disorder populations” - Introduction: “used in the ResMed S+® been shown” - Method: “Participants were fitted on the participant’s non dominant wrist with the Jawbone UP3®” - Method: “the participant subject in bed” - Method: “the three non-invasive devices was compared” - Results: “tests. 4 out of the 42 data” Should be ‘four’. - Discussion: “three consumer-grade smart devices has simultaneously

REVIEWER	Moncada-Jimenez, Jose University of Costa Rica, Human Movement Sciences Research Center (CIMOHU)
REVIEW RETURNED	18-Jan-2021

GENERAL COMMENTS	The manuscript entitled: “Accuracy of consumer-grade smart devices to measure sleep compared with polysomnography, in a sleep disorders population” attempts to investigate the convergent validity (accuracy, agreement) between portable consumer-grade devices and a gold standard used for sleep research. The focus of the topic makes this article a good fit for the BMJ Open. There are some minor concerns that need to be addressed by the authors before the manuscript is suitable for publication. My specific comments have been outlined below: 1. In general, the article is well written. 2. Page 4, lines 25 to 29: the authors should consider expand on how the different algorithms could influence sleep-related outcomes.

	3. Page 5, lines 24 to 28: the authors propose two objectives, one related to “reliability” and another related to “accuracy”. The second objective was clearly analyzed as shown in the statistical analysis section (Page 7, lines 36 to 43); however, the first objective was not clearly addressed. Thus, authors have two solve two objectives. First, reliability refers to consistent measures over time. This objective cannot be completed with the data shown in the manuscript. Second, accuracy refers to how close the consumer-grade sleep device’s results were to PSG values (i.e., convergent validity, also referred to as ‘agreement’ and ‘concordance’). Please see: Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45, 255–268. doi: 10.2307/2532051 Lin, L. I. (2000). Corrections: A note on the concordance correlation coefficient. Biometrics, 56, 324–325. doi: 10.1111/j.0006-341X.2000.00324.x I agree with Looney (2018), who suggests to report one graphical display (i.e., the Bland-Altman plots the authors reported), one scaled index (e.g., intraclass correlation coefficient with agreement definition or Lin’s Concordance Correlation Coefficient), and at least one unscaled index (e.g., root mean squared deviation, total deviation index, coverage probability, or limits of agreement) when there are two devices with no replication. Please see: Looney, M. A. (2018). Assessment of interrater and intermethod agreement in the kinesiology literature. Measurement in Physical Education and Exercise Science, 22(2), 116–128. https://doi.org/10.1080/1091367X.2017.1395742 Thus, I suggest to complete additional analysis. If possible, please compute and report Lin’s Concordance Correlation Coeffcient and compared it to the ICC you computed for the primary and secondary outcomes. Also, compute either the root mean squared deviation, total deviation index, coverage probability, or limits of agreement for the primary and secondary outcomes. These analysis will improve considerably the quality of the manuscript.

REVIEWER	Schabus , Manuel University Hospital Salzburg
REVIEW RETURNED	21-Jan-2021

GENERAL COMMENTS	In their paper, Ellender and colleagues tested the reliability of sleep reports provided by three consumer grade sleep trackers, the Jawbone UP3 as an exemplary wrist-worn device, ResMed S+ as an exemplary contactless bedside device and the Beddit sleep monitor as an exemplary mattress-based device. Independent studies comparing sleep tracking accuracy of such devices against sleep scorings derived by the gold standard PSG are of utmost importance and highly relevant for both sleep scientists and naïve end-users.

	Ellender et al. report, in addition to massive data loss due to connectivity problems, only moderate to poor reliability of the tested devices in estimating TST, WASO, SOL and SE. Overall, the paper is written well and the applied methods appear to be sound. Still, I have a few points that should be addressed before I can recommend the manuscript for publication. - p.5, l. 31 It was hypothesized that these devices would have similar accuracy in detecting TST, SOL, WASO and SE. o Given the reported results of the three devices in the introduction (e.g. Beddit … was found to have poor agreement), the above stated hypothesis is a bit surprising. Please clarify. - p.6 Polysomnographic Recording o please provide technical details about the used PSG system. o p.6., l. 26 parameters were recorded via PSG □ TST, WASO etc. were extracted and computed from the 30s sleep scoring by the expert, right? Please make this clearer. It reads like these measures were directly recorded together with the PSG data. o How is the SOL defined by the expert (Lights off until N1, until N2, until x minutes N1..?) and how is SOL defined by the different devices? If this information is not available for the consumer devices, this is an important information which should be clearly stated in the discussion section of the manuscript o Do the consumer devices all report WASO, i.e. the Wake time after sleep onset or do they report overall Wake time within the TIB period? Again, if this information is not available, mention it in the manuscript as this is an important detail against what PSG data can be compared with. o do the sleep reports by Jawbone, ResMed and Beddit also provide other information that could be compared against PSG, e.g. time spent in light, deep and/or REM sleep? Is there even a possibility to do epoch-by-epoch comparisons which would of course be of high interest? Actually this is one of the most important points in my view as it would be highly desirable to do an epoch-by-epoch comparison rather a comparison of average whole night values where there is little control if not alone sex and age provided to the devices biases the consumer device output. Please check if such an analysis can be done and at least added this point to the discussion section. - p.7. Statistical analyses o The authors state that normality was assessed by SW tests. Please also clearly state that the normality assumptions were not violated and add the SW test statistics. o p.9, l. 17; spelling/grammar mistake: No significant were found between PSG and the test devices were mediated by gender, . □ Please rephrase that sentence and explain these control analyses in more detail. o p.9, line 24. ..however not usually in the same room or on the same patient □ Please provide more information and descriptive data for this point. - Bland Altman Plots o I suggest to provide multi panel plots (e.g. 1x3 for each metric) using the same scales to be able to better compare the results of the devices.

	o please provide Bland Altman plots for SOL and WASO, too (if not possible in the main text in the supplements). - Discussion o Are there studies or can you please speculate about the accuracy (drop?) of contactless bedside and mattress-based devices when the subject lies in bed with a partner/child/animal? I believe this is the ecologically valid condition in which people usually will sleep at home. At least such a limitation should be mentioned and discussed. Minor: p.5., l.10, spelling error. The sensor technology used in the ResMed S+® been

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 –
• Methods has been amended to address lights on/off sync and that all studies were scored by a blinded RPSGT. Version of device interface have been included.
• Limits of Agreement and bias has been included in an updated Table 3
• Typographical errors have been amended with thanks

Reviewer 2 –
• Manuscript has been amended to clarify specific statistical tests addressing agreement and reliability.
• Further statistical support has been obtained to tighten the reporting of limits of agreement and bias with an updated Table 3. This provides a similar statistical outcome to that suggested by reviewer 2.

Reviewer 3 –
• Methods have been amended to clarify PSG recording, which was undertaken as per AASM standards.
• Multi-panel Bland-Altman's will be submitted as a supplementary figure to clarify.

VERSION 2 – REVIEW

REVIEWER	Scott, Hannah Flinders University
REVIEW RETURNED	17-May-2021

GENERAL COMMENTS	No further comments, apart from adding lines of best fit to Bland-Altman Plots and SDs/CIs to Means presented in the Abstract, if another round of reviews is needed.

REVIEWER	Schabus , Manuel University Hospital Salzburg
REVIEW RETURNED	07-Jun-2021

GENERAL COMMENTS	The study conducted by Ellender and colleagues aims at evaluating the performance of different consumer-grade smart devices to measure sleep-related variables. Results indicate that there is poor to moderate agreement between the devices and PSG, with the wrist actigraph (Jawbone UP3) performing the best.

	In addition, the authors also reported a high device synchronisation failure, especially for the Beddit device and leading to considerable data loss. The main contribution of this study primarily lies in investigating these devices' performance in a sample with sleep related disorders. The proposed study is methodologically acceptable. A sample size of 44 adult patients has been acquired, devices' performance has been compared to the gold standard (PSG), and accuracy has been estimated using appropriate statistical methods and techniques (i.e., intra-class correlations, Bland-Altman plots). However, the absence of an epoch-by-epoch (for each 30sec) analysis strongly undermines the impact that this study could have. It is well known that for averaged estimates like total sleep time or wake after sleep onset accuracy agreement is much easier to achieve. Partly this may be because algorithms are aware what are plausible values for different age groups and therefore these devices usually deviate most strongly for unexpected good or bad nights, but not the average night. It is important here also to note that the data have been collected a long time ago (2015-2016) and thus, to date, some of the devices are not available (e.g., to my knowledge Jawbone is not available since 2017!), or they might at least have updated sensors and software. This needs to be clearly indicated as limitation in abstract and discussion; that is, that the paper should be more seen as a proof of principal investigation than a timely evaluation of sensors for sleep staging. In addition, the manuscript is poorly written, the graphs have been carelessly produced, and the discussion of the results is insufficient and partly inadequate. I thus strongly encourage the authors to address at least the following points to make it a reasonable contribution to the field: Major: ● The authors need to revise the manuscript for readability and clear writing. For example, lines 52-53 page 9: “Generally, PSG measured TST 157mins below or 100mins above Jawbone, however, data were closer between 300-400min.” As the PSG is the gold standard in sleep research Jawbone either overestimates or underestimates a given variable. The current writing leads to confusion. Please also double check the direction of the effects and if the y-axis is labelled correctly. I would be very surprised if the consumer device is not overestimating sleep in poor/short (low TST) nights and underestimating sleep in exceptional long (a lot of TST) nights. ● Since the data have been collected in 2015-2016, it is mandatory that the authors add in the discussion that some of the devices might have been updated, and check to what extent at the present day some of the devices have improved their sensors, and software (e.g., Bluetooth connection). Jawbone for instance is not available since 2017 to my knowledge. The authors should clearly acknowledge this also in the abstract, in order to make sure that it is clear to the reader how the results of this study should be interpreted. The way it currently reads it certainly implies that the Jawborne is an actual consumer devices that should be checked out.

	● The authors need to discuss the extent to which they expect to find an agreement between non-brain sleep classifications and PSG in the first place. The transition from wake to sleep is a gradual process and defined by the subjects' current brain state and not behavioural state. Taking this under consideration, what level of agreement do sleep scientists expect to find between peripheral activity and brain measurements? The authors should discuss the matter as this is an important and relevant theoretical point. It is a topic discussed heavily in sleep research and there are plenty of papers comparing PSG to actigraph data. ● In addition it is highly needed to specify more clearly what exactly the tested devices record. That is, do they rely on pure activity data for the sleep parameters they supply or is it a combination with some sort of heart rate data (PPG, Interbeat intervals [IBI], variance of IBIs, … ?). Only that way the reader can judge the results reasonably. ● The visualization of the results is poor for the moment: please improve the quality of all figures, add metrics on y axes (e.g., minutes), and define the “MD” label in figure 2, for example (Alternatively call it “bias” as in Tables, but in any case it should be consistent). Put Bland-Altman figures in one plot with corresponding titles ● I assume the figure in page 27 is not a figure that is meant to be included in the paper-since it is completely missing a figure caption. ● The authors should better relate their findings to the corresponding figures and link it. ● The authors should also discuss the results of the “Beddit device”. ● The authors might consider moving the demographic information in the methods section, as this is not really a result of the experimental protocol. ● It is also not clear why 12 participants from the initial sample were excluded. The authors should include this detail in the methods section. Please clarify what was the final sample for each and every device. ● For the cases of ResMed and the Beddit device, some extreme differences for TST are observed. Have the authors checked whether the data with these extreme differences (e.g. 200min off in total sleep time estimation) come from patients with specific sleep disorders? That is, are this for example people with apnea or adipositas or random subject where these devices are so far off? Comment on it in the paper. ● If possible colour code the data based on the biggest patient groups (e.g., insomnia, apnea, etc.). This might help to better evaluate the tested devices and the outcome of the bland-altman analysis.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1 Dr. Hannah Scott, Flinders University

No further comments, apart from adding lines of best fit to Bland-Altman Plots and SDs/CIs to Means presented in the Abstract, if another round of reviews is needed.

Author response:

• Lines of best fit have been added to figures and repeat images uploaded. Thank you

Reviewer :2 no comments

Reviewer: 3 Prof. Manuel Schabus , University Hospital Salzburg

● The authors need to revise the manuscript for readability and clear writing. For example, lines 52-53 page 9: “Generally, PSG measured TST 157mins below or 100mins above Jawbone, however, data were closer between 300-400min.” As the PSG is the gold standard in sleep research Jawbone either overestimates or underestimates a given variable. The current writing leads to confusion. Please also double check the direction of the effects and if the y-axis is labelled correctly. I would be very surprised if the consumer device is not overestimating sleep in poor/short (low TST) nights and underestimating sleep in exceptional long (a lot of TST) nights.

● Since the data have been collected in 2015-2016, it is mandatory that the authors add in the discussion that some of the devices might have been updated, and check to what extent at the present day some of the devices have improved their sensors, and software (e.g., Bluetooth connection). Jawbone for instance is not available since 2017 to my knowledge. The authors should clearly acknowledge this also in the abstract, in order to make sure that it is clear to the reader how the results of this study should be interpreted. The way it currently reads it certainly implies that the Jawborne is an actual consumer devices that should be checked out.

● The authors need to discuss the extent to which they expect to find an agreement between non-brain sleep classifications and PSG in the first place. The transition from wake to sleep is a gradual process and defined by the subjects' current brain state and not behavioural state. Taking this under consideration, what level of agreement do sleep scientists expect to find between peripheral activity and brain measurements? The authors should discuss the matter as this is an important and relevant theoretical point. It is a topic discussed heavily in sleep research and there are plenty of papers comparing PSG to actigraph data.

● In addition it is highly needed to specify more clearly what exactly the tested devices record. That is, do they rely on pure activity data for the sleep parameters they supply or is it a combination with some sort of heart rate data (PPG, Interbeat intervals [IBI], variance of IBIs, … ?). Only that way the reader can judge the results reasonably.

● The visualization of the results is poor for the moment: please improve the quality of all figures, add metrics on y axes (e.g., minutes), and define the “MD” label in figure 2, for example (Alternatively call it “bias” as in Tables, but in any case it should be consistent). Put Bland-Altman figures in one plot with corresponding titles

● I assume the figure in page 27 is not a figure that is meant to be included in the paper-since it is completely missing a figure caption.

● The authors should better relate their findings to the corresponding figures and link it.

● The authors should also discuss the results of the “Beddit device”.

● The authors might consider moving the demographic information in the methods section, as this is not really a result of the experimental protocol.

● It is also not clear why 12 participants from the initial sample were excluded. The authors should include this detail in the methods section. Please clarify what was the final sample for each and every device.

● For the cases of ResMed and the Beddit device, some extreme differences for TST are observed. Have the authors checked whether the data with these extreme differences (e.g. 200min off in total sleep time estimation) come from patients with specific sleep disorders?

● If possible colour code the data based on the biggest patient groups (e.g., insomnia, apnea, etc.). This might help to better evaluate the tested devices and the outcome of the bland-altman analysis.

Author’s response

• Results section has been updated to clarify communication

• Plots were labelled and interpreted correctly in the first draft. However, realizing the confusion, we have changed the calculation of mean difference (new device-PSG) instead of calculating the difference by subtracting outcome measured by PSG from outcome measured by the new device. Now the negative mean difference indicates under-estimation to ease interpretation. The changes are reflected in mean difference and upper and lower agreement limits. The explanation in the results section will further clarity the graphs and link the figures better to the main document, addressing another concern raised.

• In our findings two of the devices did not overestimate sleep. Whilst this has not been found for the Beddit device previously, the previous studies only had 10 people and were healthy controls.

• Issue regarding the age of these devices has been amended in the conclusions of the abstract. Further detail has been added to the limitations section of the discussion with mention of the outcomes of all three devices over time.

• With regards to discussion of non-brain sleep monitoring, this is beyond the scope of this article. There is extensive discussion in the literature, role and clinical utility for non-sleep monitors well established from the 2018 American Academy of Sleep Medicine position statement in the use of actigraphy, a comparable concept. Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, Carden KA. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med. 2018;14(7):1209–1230.

• The technology for each of the devices is described in brief in the methods. More detail here is beyond the scope for this article, which is targeted at clinicians.

• The 12 patients not included in the analysis were due to device synchronisation failure – the results of this is outlined in the results section “Consumer-grade recording failure”

• Figure axis and labelling improved with section for interpretation added in results section

• Whilst a further sub-group analysis of the outliers would be of interest, the authors do not feel this is within the scope of this small sample size. Should larger samples be undertaken, this would be of interest.

We appreciate the time undertaken for this clinically focused paper to undergo peer review and that given there are two favourable reviewer comments, hope this manuscript can move forward.

VERSION 3 – REVIEW

REVIEWER	Scott, Hannah Flinders University
REVIEW RETURNED	23-Aug-2021

GENERAL COMMENTS	I commend the authors for addressing all reviewer comments in the past two rounds of review. This paper makes a strong contribution to the literature. The lack of epoch-by-epoch analyses is certainly a study limitation, and is appropriately acknowledged as such in the manuscript. The analyses that have been conducted, however, are appropriate for drawing the conclusions that the authors make in the Discussion section. Therefore, I do not think that the lack of epoch-by-epoch analyses is a 'deal-breaker' for this manuscript. It is my opinion that the manuscript should thus be accepted as is.

REVIEWER	Schabus , Manuel University Hospital Salzburg
REVIEW RETURNED	20-Aug-2021

GENERAL COMMENTS	After addressing the relevant points, I think that the quality of the current clinically-focused manuscript has been overall improved. The readability and clarity of the results section have been also substantially improved. The plots are now linked to the results adequately, and the overall figure quality, with axis clarification and more detailed captions, has benefited by the author's implementations. The authors have added relevant parts in the discussion and acknowledged one of the main limitations of the current work: the devices tested here are outdated, and some, in fact, are not available in the market today. Despite that, the study serves as principle study on a large clinical sample and thus its results are relevant to clinical work. However it is important, I argue, that the authors compare their results with more recent studies that use, if not the same, very similar devices (as in the case of ResMed, for instance), so that the impact of their study reaches current relevant clinical practices.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1 Dr. Hannah Scott, Flinders University

Accepted as is.
Author response: Thank you

Reviewer 3. Prof. Manuel Schabus , University Hospital Salzburg
However it is important, I argue, that the authors compare their results with more recent studies that use, if not the same, very similar devices (as in the case of ResMed, for instance), so that the impact of their study reaches current relevant clinical practices.
Author response:
• We have referenced in the introduction and discussion a recent similar comparison undertaken by Chinoy et al (2021) in which the ResMed S+ was compared with PSG, in a health normal population. The sample size of this study is smaller than current paper there was stronger agreement between gold standard and the ResMed S + and the next generation device. We suspect this difference reflects the study population difference – a clinic versus health normal population. Interestingly this group also found device failure synchronisation to be an issue in their series.